RHIC Computing Facility

Eric Lançon

August 23, 2016

BROOKHAVEN NATIONAL LABORATORY

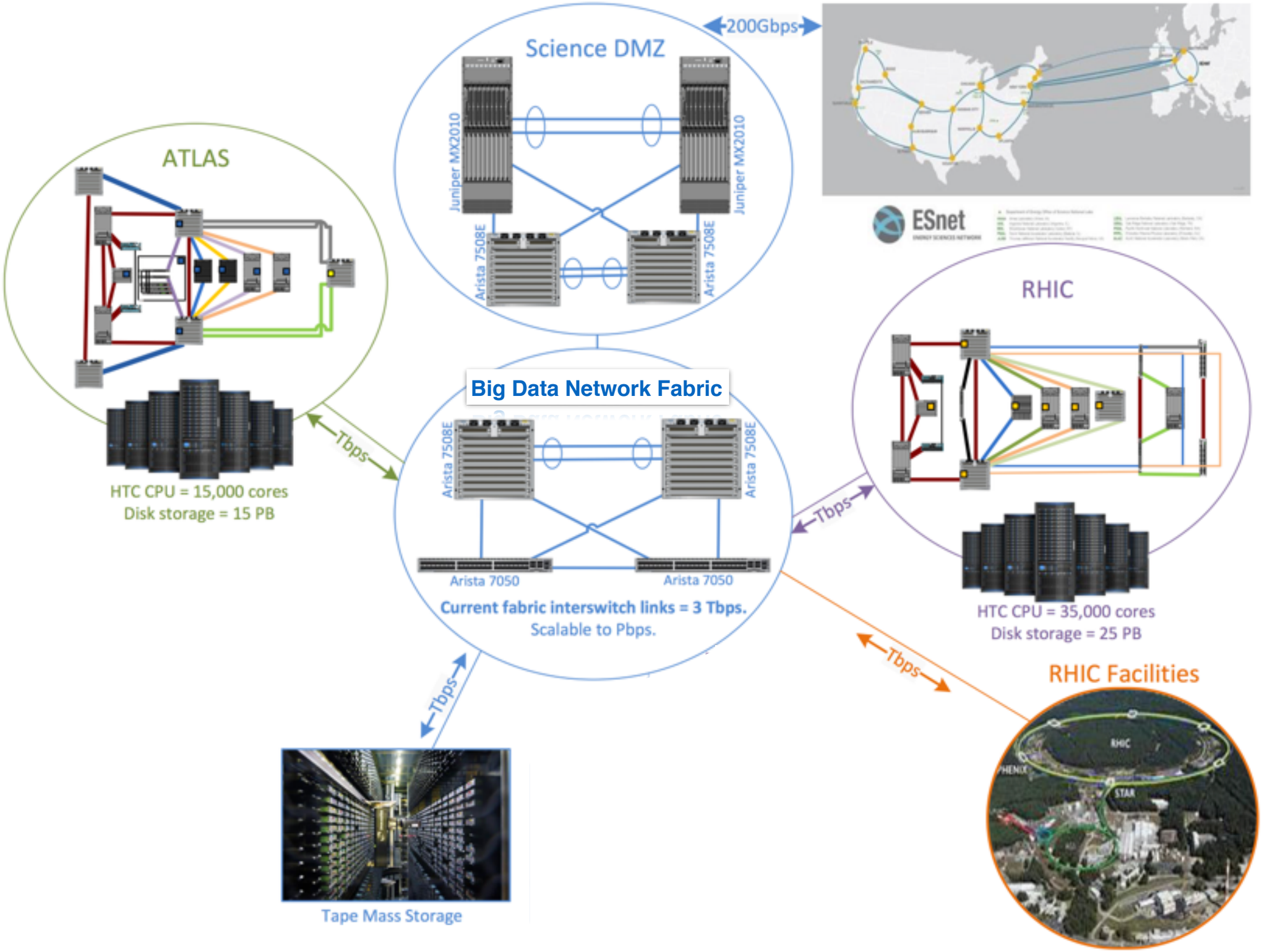
a passion for discovery

Office of Science
U.S. DEPARTMENT OF ENERGY

Outline

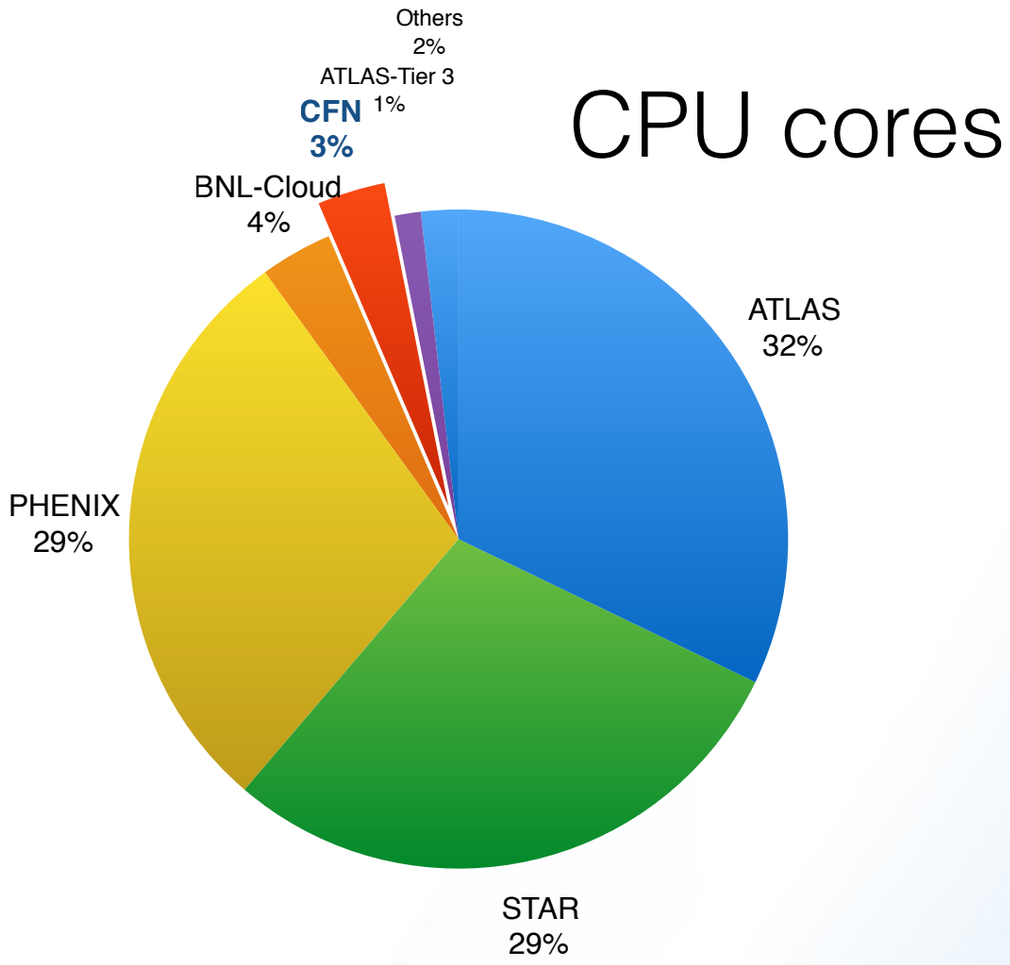
- Status of RCF, synergies with ATLAS Tier-1
- Performance in recent RHIC runs
- Future technological and data challenges
- Synergies with BNL Computing Initiative
- B725 infrastructure project

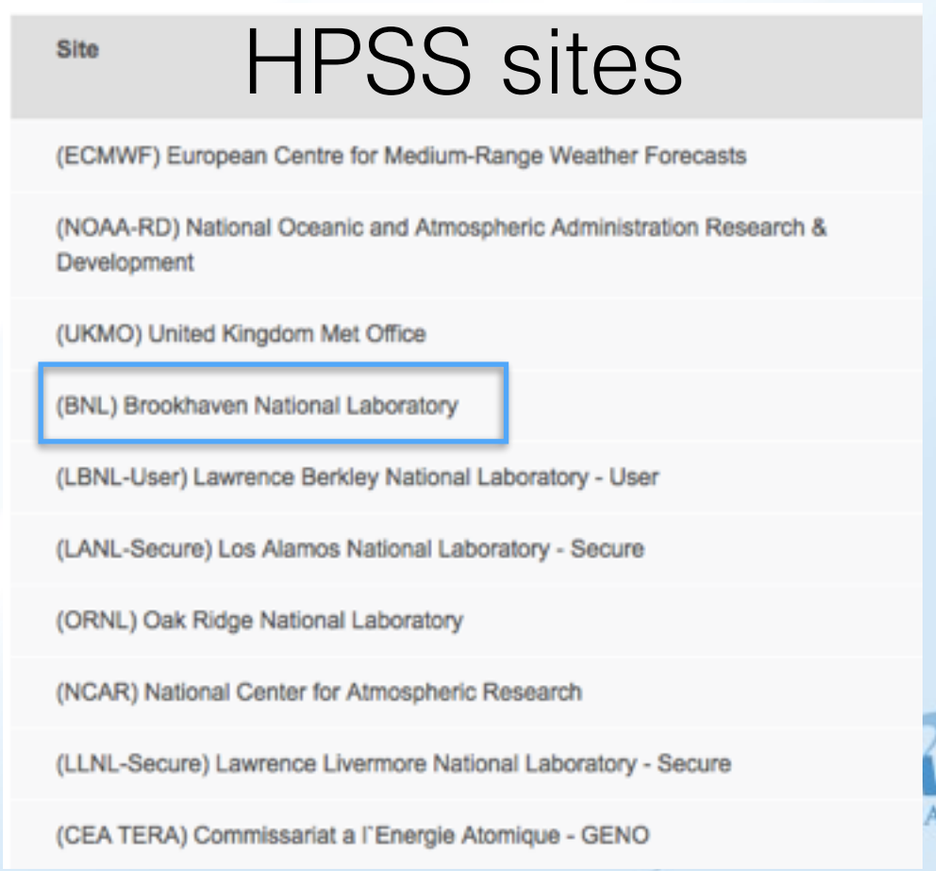
RCF today


ATLAS
HTC CPU = 15,000 cores
Disk storage = 15 PB
Tbps
Science DMZ
Juniper MX2010
Juniper MX2010
Arista 7508E
Arista 7508E
200Gbps
ESnet
Big Data Network Fabric
Arista 7508E
Arista 7508E
Arista 7050
Arista 7050
Current fabric interswitch links = 3 Tbps.
Scalable to Pbps.
Tbps
Tbps
Tbps
RHIC
HTC CPU = 35,000 cores
Disk storage = 25 PB
RHIC Facilities
RHIC
PHENIX
STAR
Tape Mass Storage

Capacities as of today

- **55k CPU cores**
  - **3%** HPC of capacity, will increase in the next months
- **~45 PB of disk storage**
  - of various technologies
- **~80 PB of tape storage**
  - 4th HPSS site worldwide
  - first within the US(1)





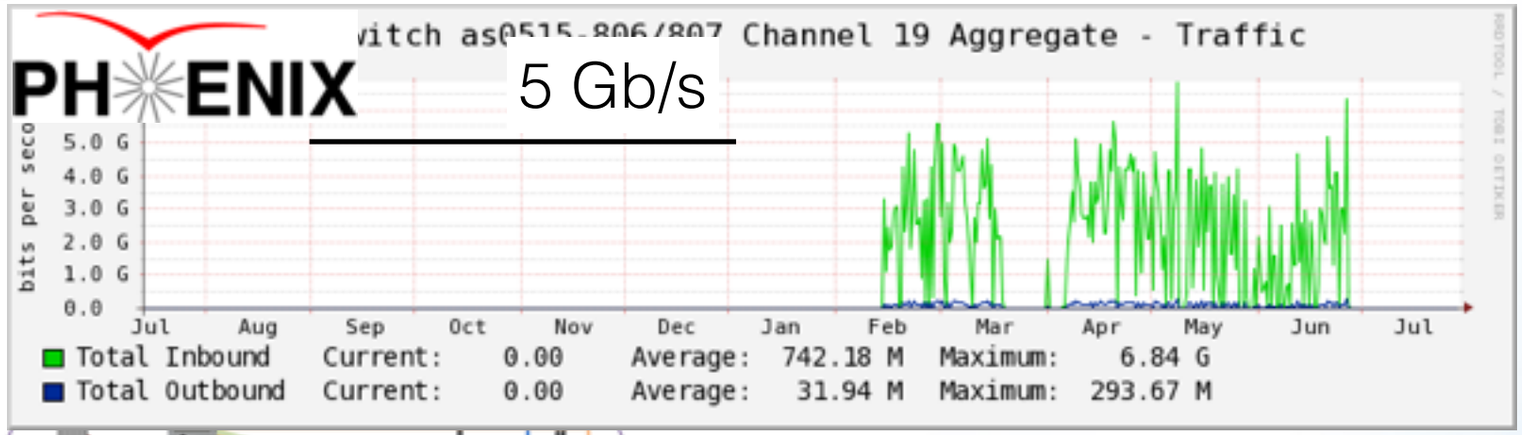
(1) http://www.hpss-collaboration.org/learn_who_petabyte_data.shtml

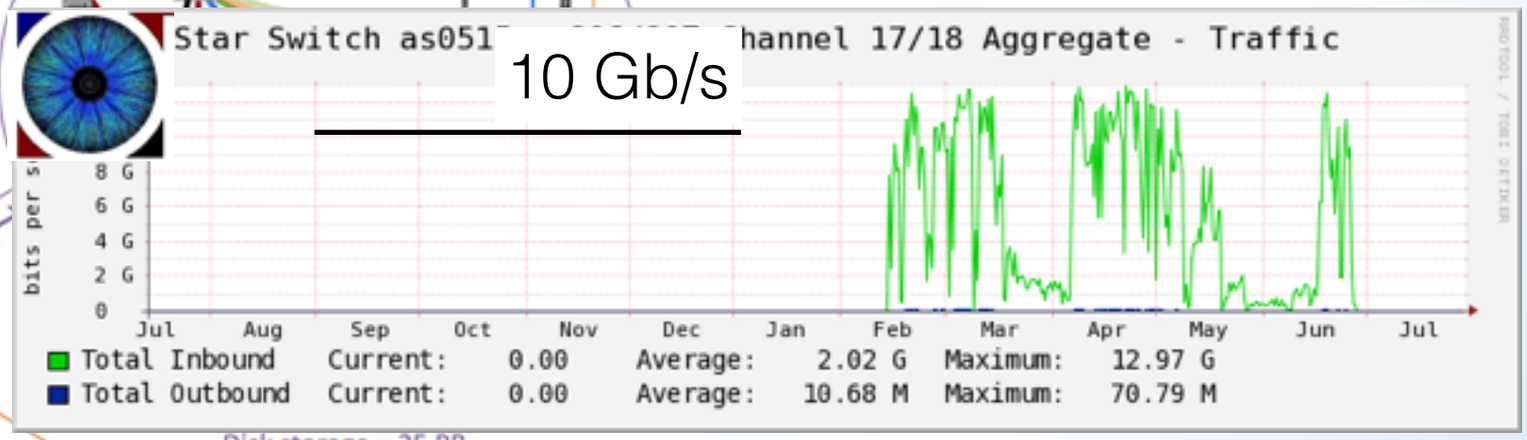
Status of RCF

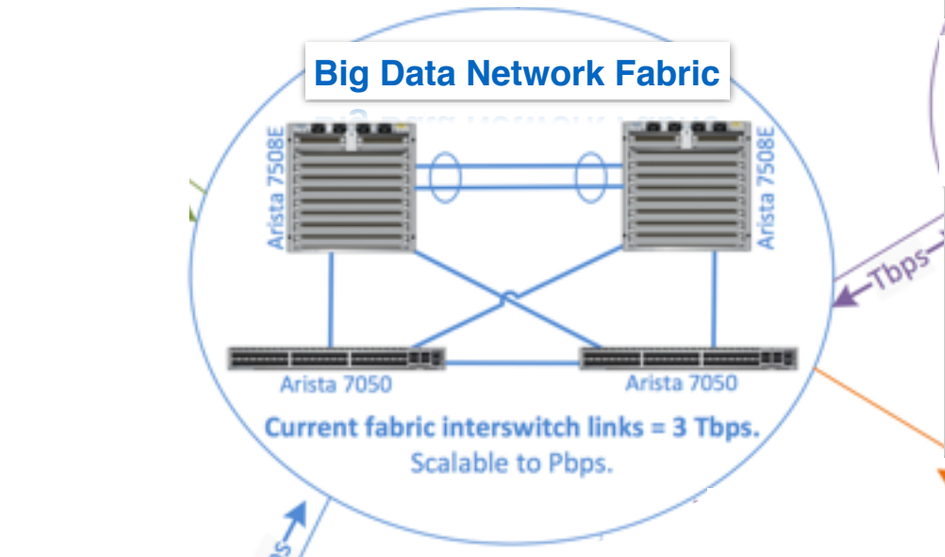
- RCF performed well during 2016 run
- Resources are ~fully utilised
- Hardware (CPU) is getting old, migration to new tape generation needed (space in HPSS)
- Increase of resources needed in the next years

Performance in 2016

- No issue in data transfer from experiments to facility

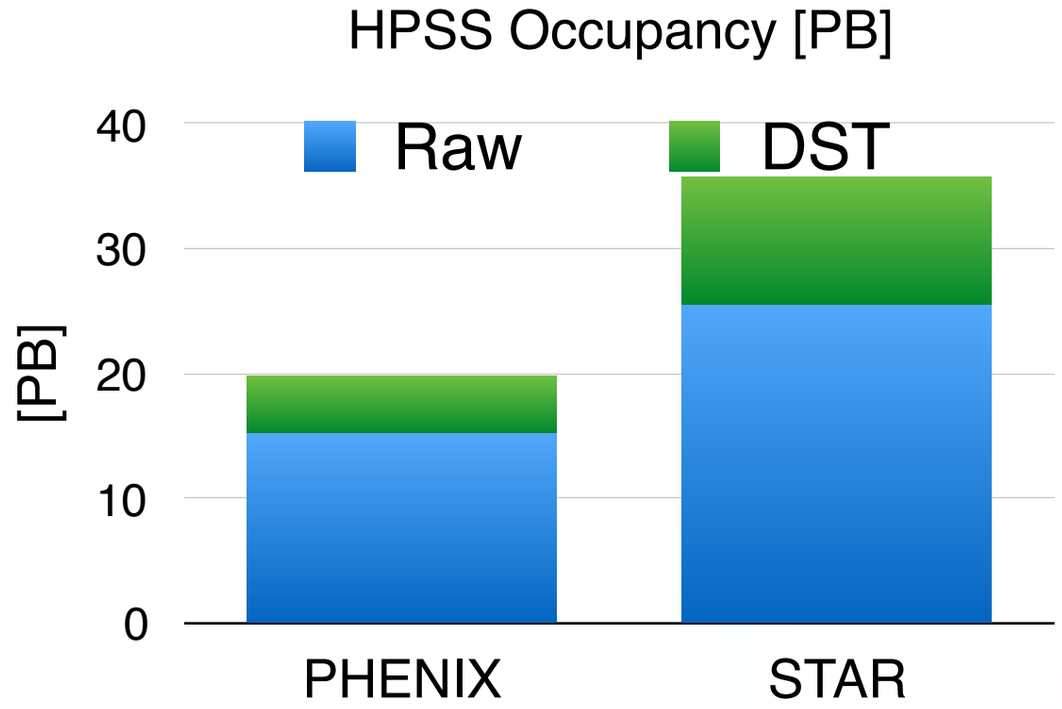


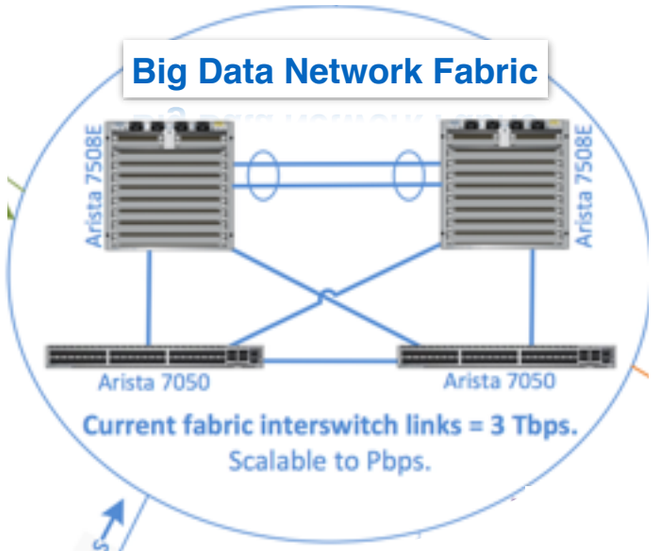




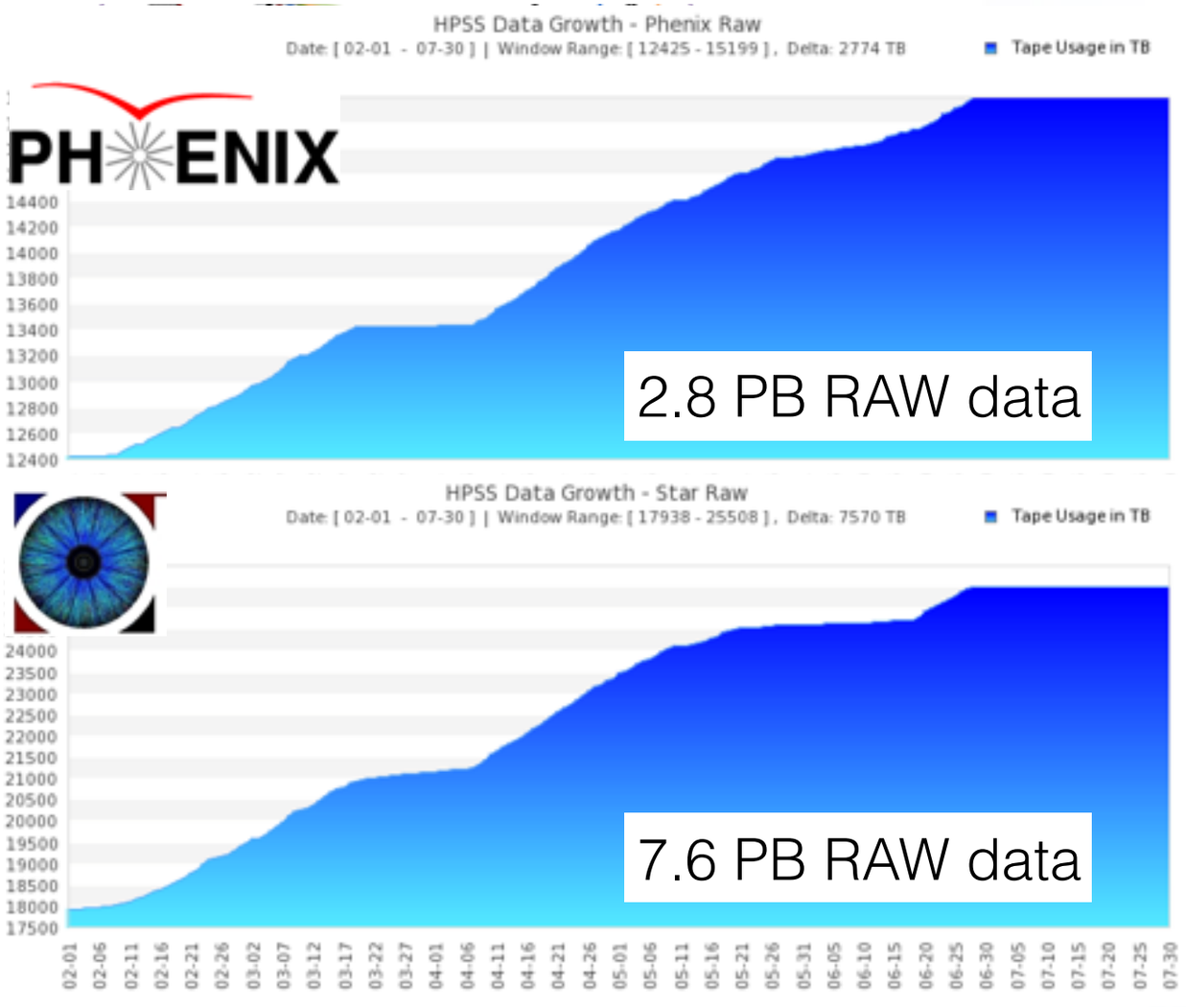
Performance in 2016

- No issue in writing RAW data to tape



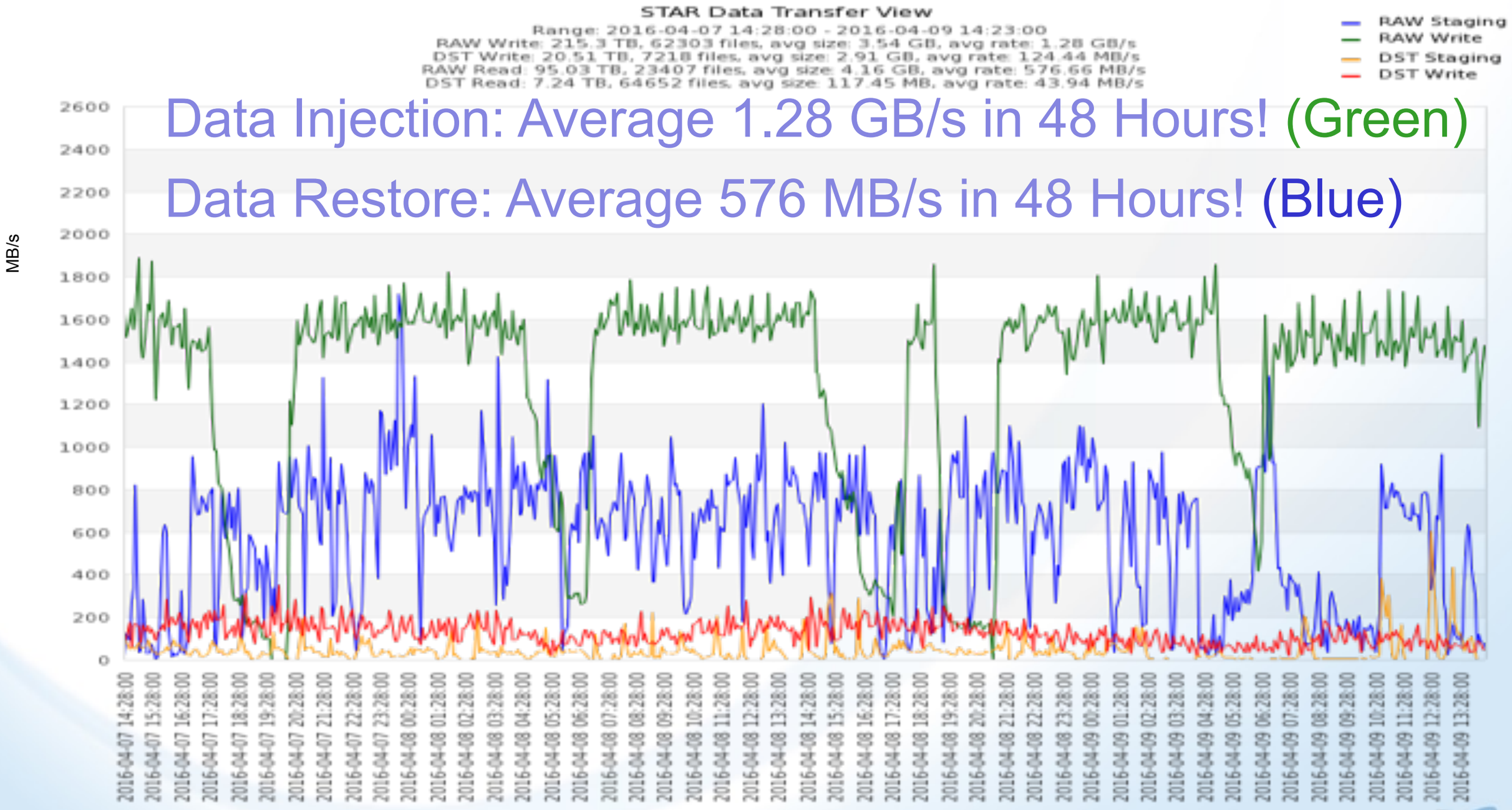


Tape Mass Storage



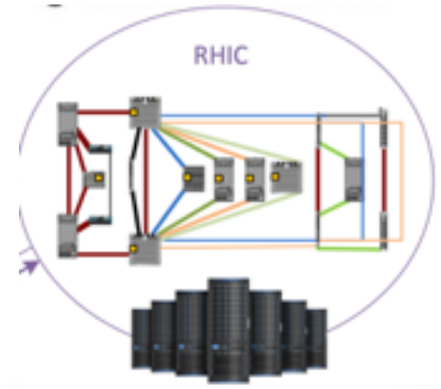
High Throughput Parallel Archiving

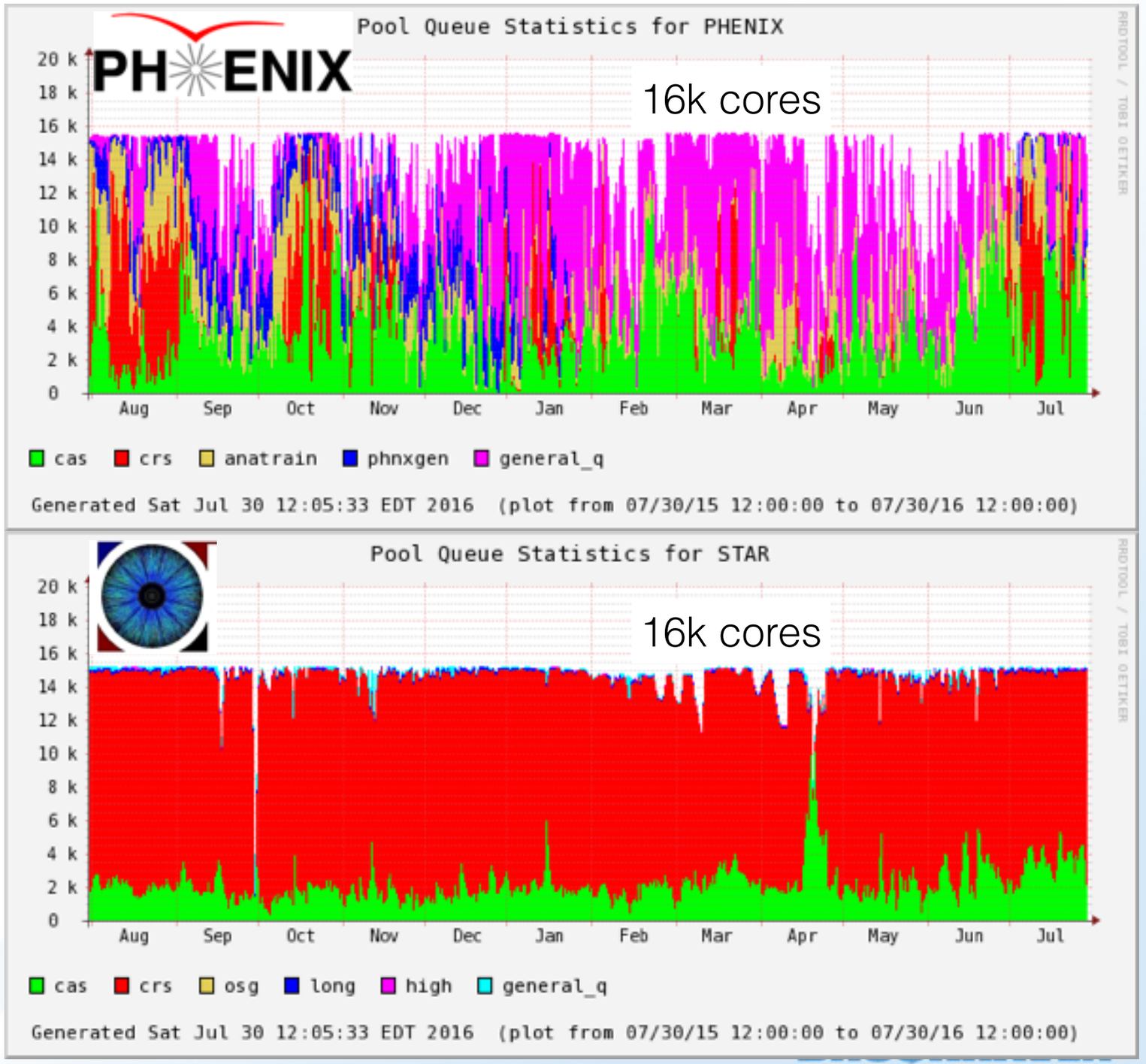
RHIC RUN 16 - STAR



Performance in 2016

- Two distinct computing farms of equal size, one for PHENIX, one for STAR
- Storage distributed on computing nodes
  - Reconstruction jobs of experience A cannot run on farm of experience B
- STAR farm almost continuously saturated while PHENIX farm is not

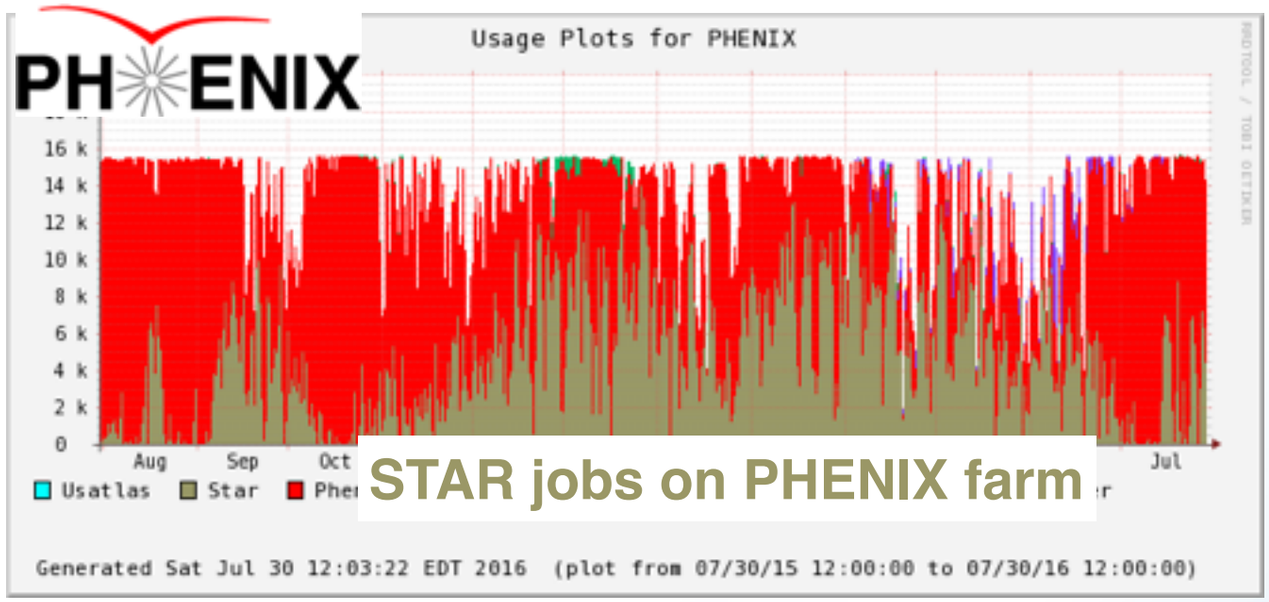


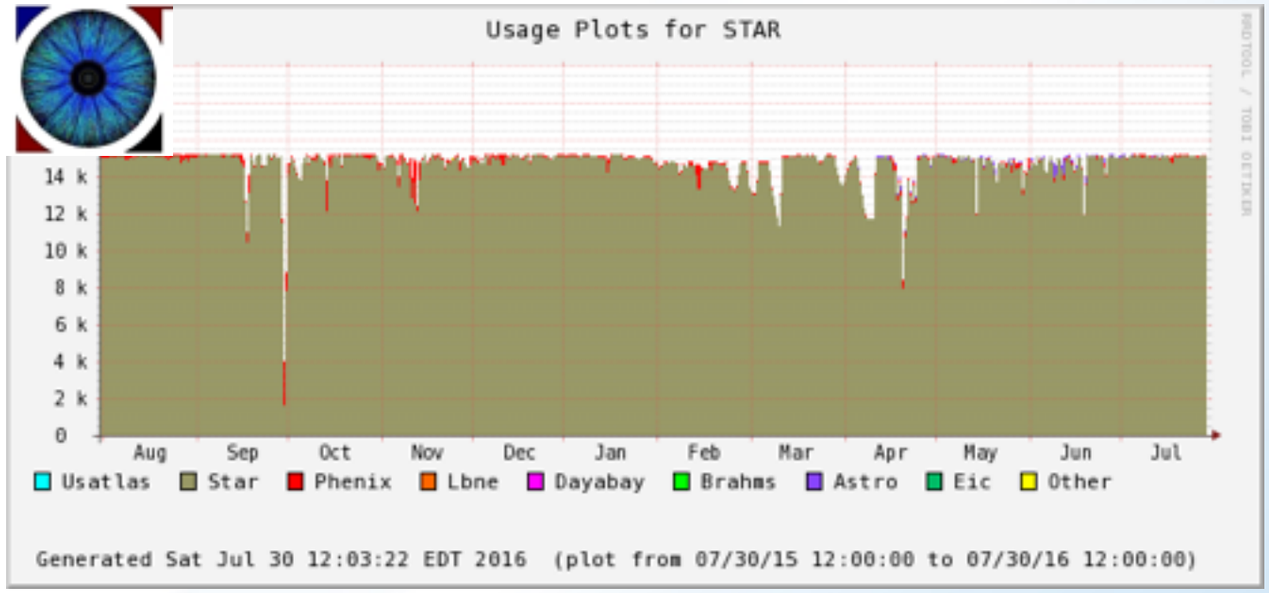


CPU usage of the farms

- PHENIX farm used by STAR **analysis** jobs when no PHENIX activity
  - Optimisation of batch system (Condor) performed by RCF,
  - STAR analysis workflow optimisation to be done (too long jobs)

- Lesson for the future
  - Computing models (workflow management, data organisation,…) and technological choices (storage, CPU,…) of experiments should not be too different in order to benefit from a global pool of resources



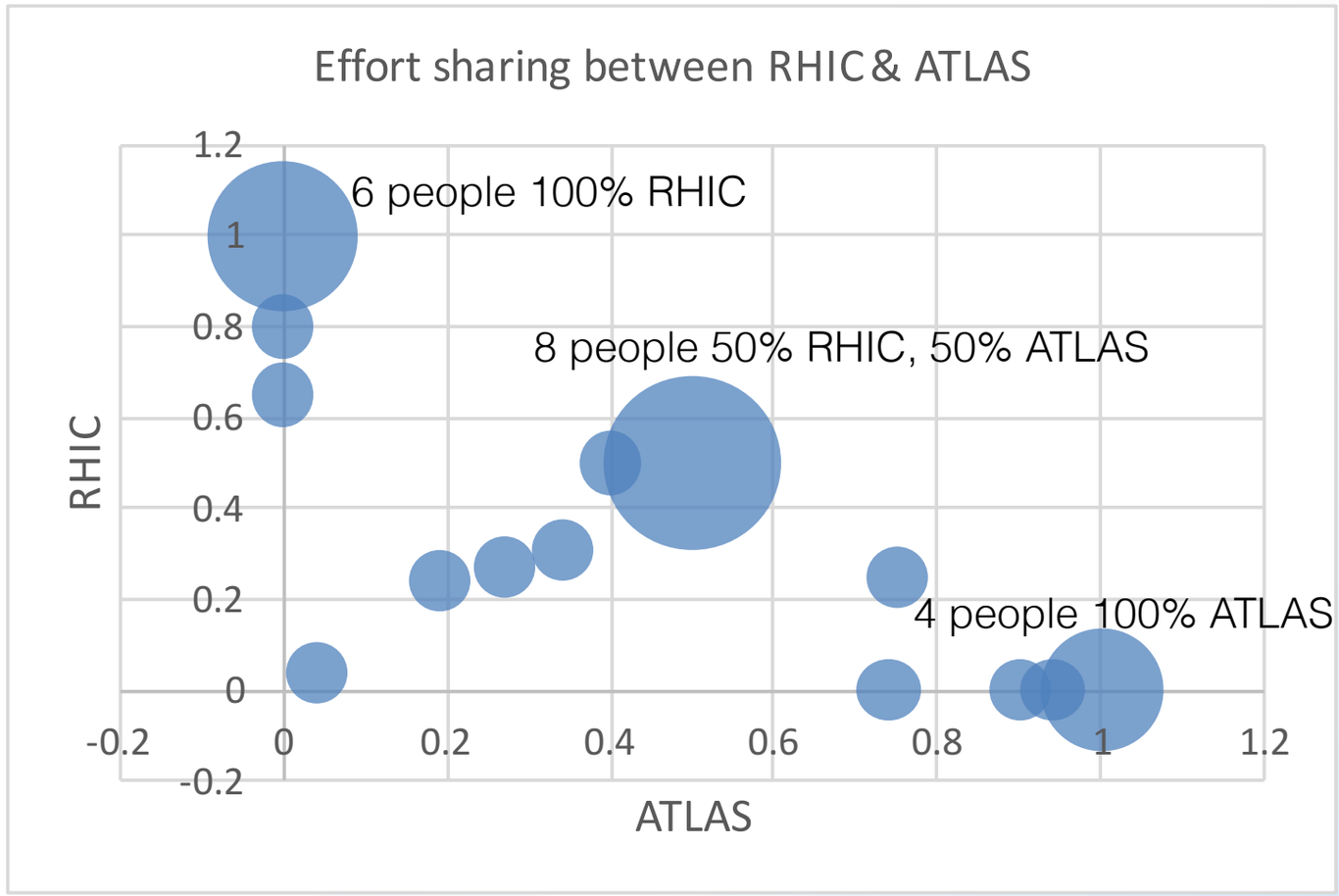


Synergy with ATLAS Tier-1

- Economy of scale (operation, purchase,…)
- Common procedures and configurations (resilience)
- Common tools (batch system, storage, network)
- Expertise from RCF benefits to ATLAS (and vis versa)
- Access to Grid and cloud computing expertise developed in ATLAS
- …

Synergy with ATLAS Tier-1

- 13.6 FTE for RHIC
  - Support from ITD included
  - 6 people are 100% RHIC (storage, infrastructure, user support,…)
  - 8 people 50/50 (batch, system administration & configuration,…)
- About the right size of effort provided new RHIC experiments do not develop complex computing models

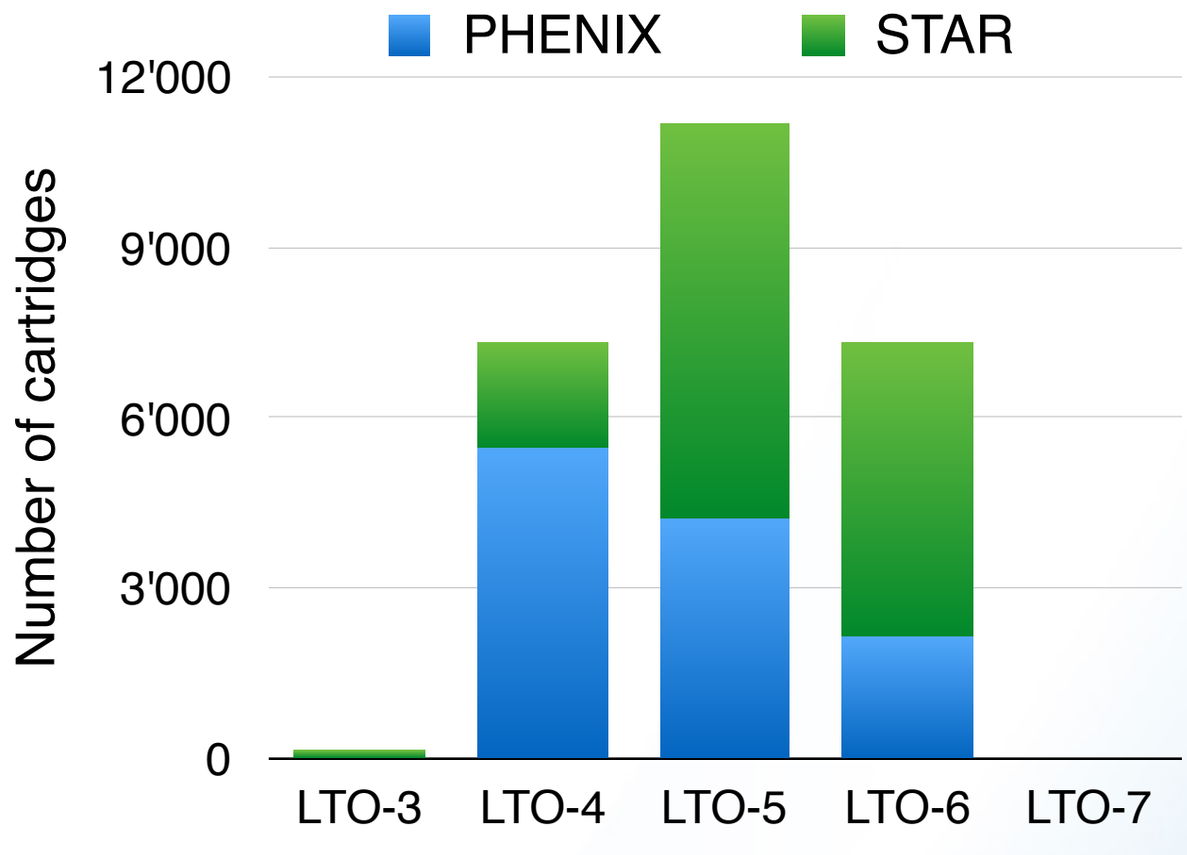


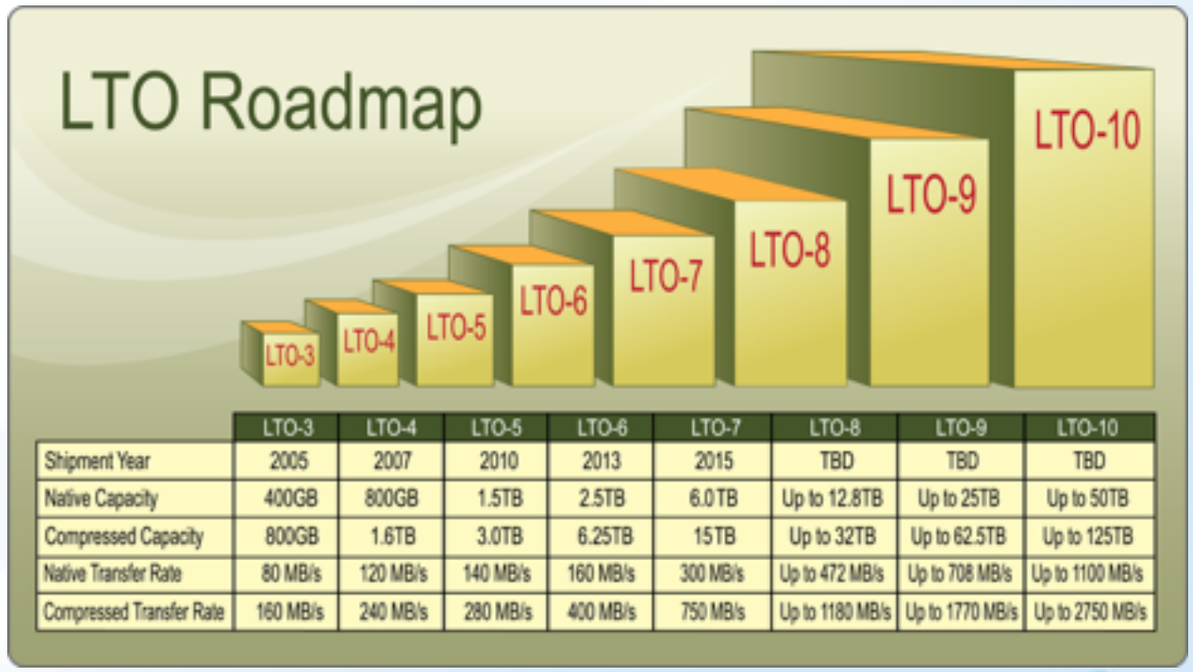
Future technological and data challenges

- **Future of computing is multi-core**
  - New hardware are multi-core 16, 32, 64,…. with less and less memory per core
  - Could software of RHIC experiments be multi-core?
  - Is it worst the effort for existing experiments?
- **Object store technology**
  - ATLAS will migrate to Ceph (2-5 years)
  - To be considered for sPHENIX and eRHIC
- **RHIC hardware is getting old**, ~25% older than 5 years
- **Tape technology**
  - 2 generations behind in tape technology
  - Only one copy of RAW date on tape
- **Data preservation**
  - Access to data and software after data taking

Tape migration

- Need to migrate archived data to new tape technology (LTO-7)
  - ~7 more capacity / tape
  - ~3 time faster
- LTO-7 tape drives cannot read LTO-4 and older types
  - Data on LTO-4 copied onto LTO-7
- 2 copies of RAW data will be made in the migration process
  - Today 1 copy of RAW data

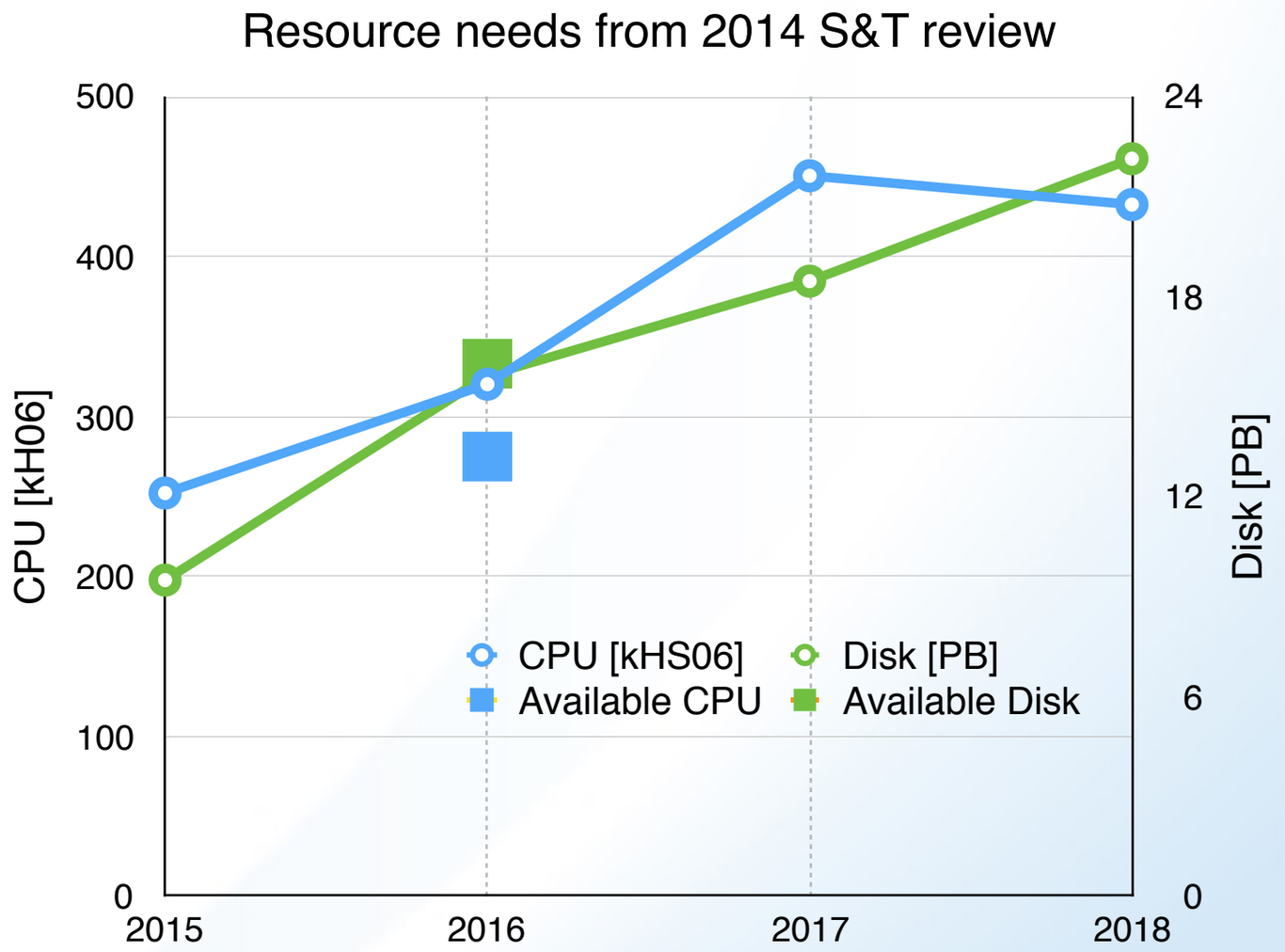




	LTO-3	LTO-4	LTO-5	LTO-6	LTO-7	LTO-8	LTO-9	LTO-10
Shipment Year	2005	2007	2010	2013	2015	TBD	TBD	TBD
Native Capacity	400GB	800GB	1.5TB	2.5TB	6.0TB	Up to 12.8TB	Up to 25TB	Up to 50TB
Compressed Capacity	800GB	1.6TB	3.0TB	6.25TB	15TB	Up to 32TB	Up to 62.5TB	Up to 125TB
Native Transfer Rate	80 MB/s	120 MB/s	140 MB/s	160 MB/s	300 MB/s	Up to 472 MB/s	Up to 708 MB/s	Up to 1100 MB/s
Compressed Transfer Rate	160 MB/s	240 MB/s	280 MB/s	400 MB/s	750 MB/s	Up to 1180 MB/s	Up to 1770 MB/s	Up to 2750 MB/s

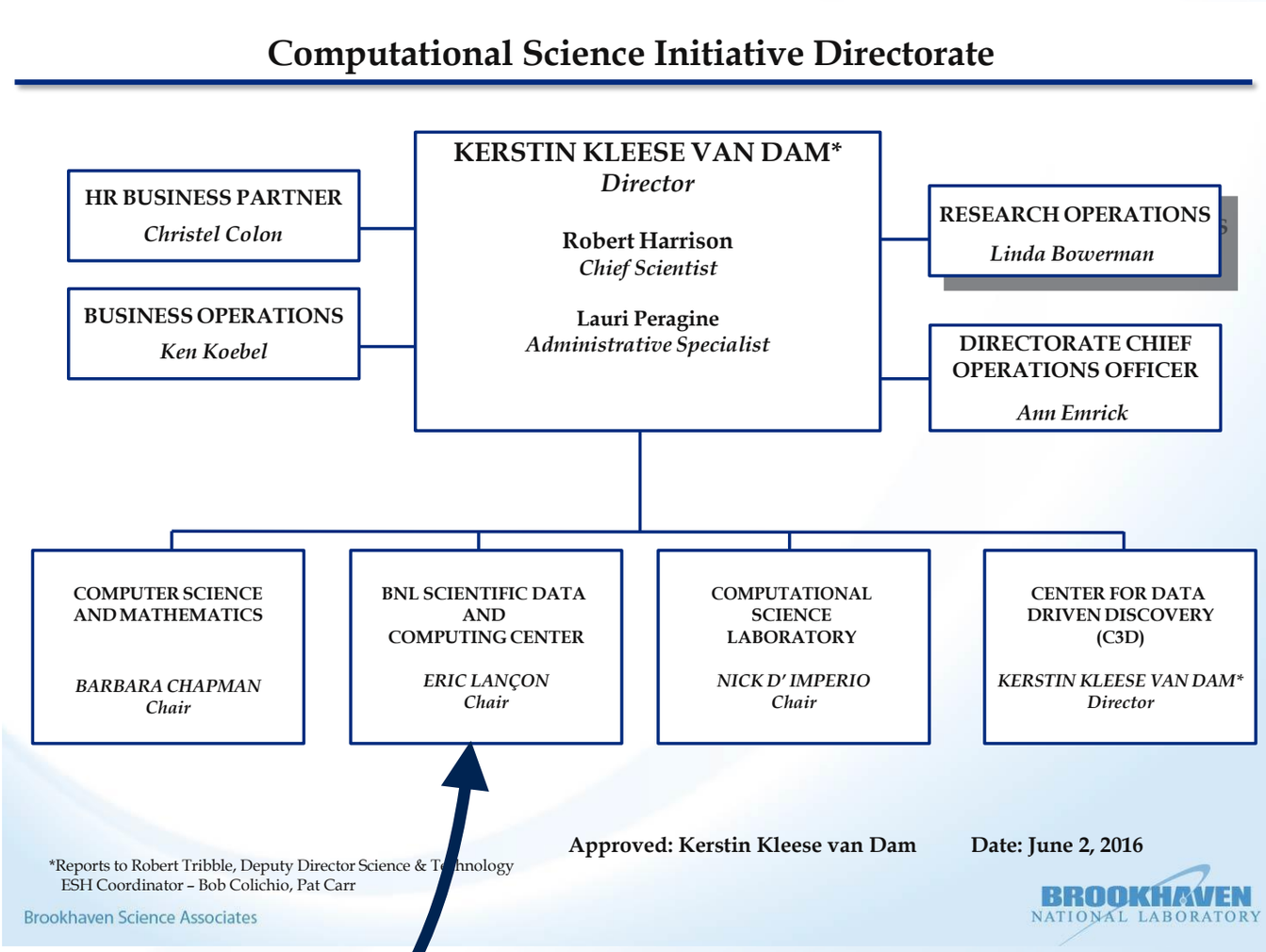
CPU & Disk resources for next years

- Today's resources just match anticipated needs from 2014 S&T review
- 25% of capacity is older than 5 years and need to be replaced
- Projected 2017 needs (including replacement) ~1.6 what is currently installed
- Projection did not include running in 2017
- **Real 2017 needs ~1.8 current capacity**



Computational Science Initiative : CSI

- **CSI** *: Integrating data-intensive science expertise and investments across the Laboratory to tackle "big data" challenges*
  - Leverage investments across multiple programs
  - Patterns : universities (Columbia, Cornell, New York University, Stony Brook, and Yale) and companies including IBM Research.
- **SDCC:** Scientific Data and Computing Center of CSI



RHIC and ATLAS Computing Facility operates SDCC

- **SDCC** is the computing center of CSI, BNL's Computational Science Initiative
- It is operated by **RACF**
- It includes components from
  - Laboratory's Institutional Cluster
  - CFN (Center for Functional Nano-materials)
  - Atmospheric Radiation Measurement
  - USQCD
  - …

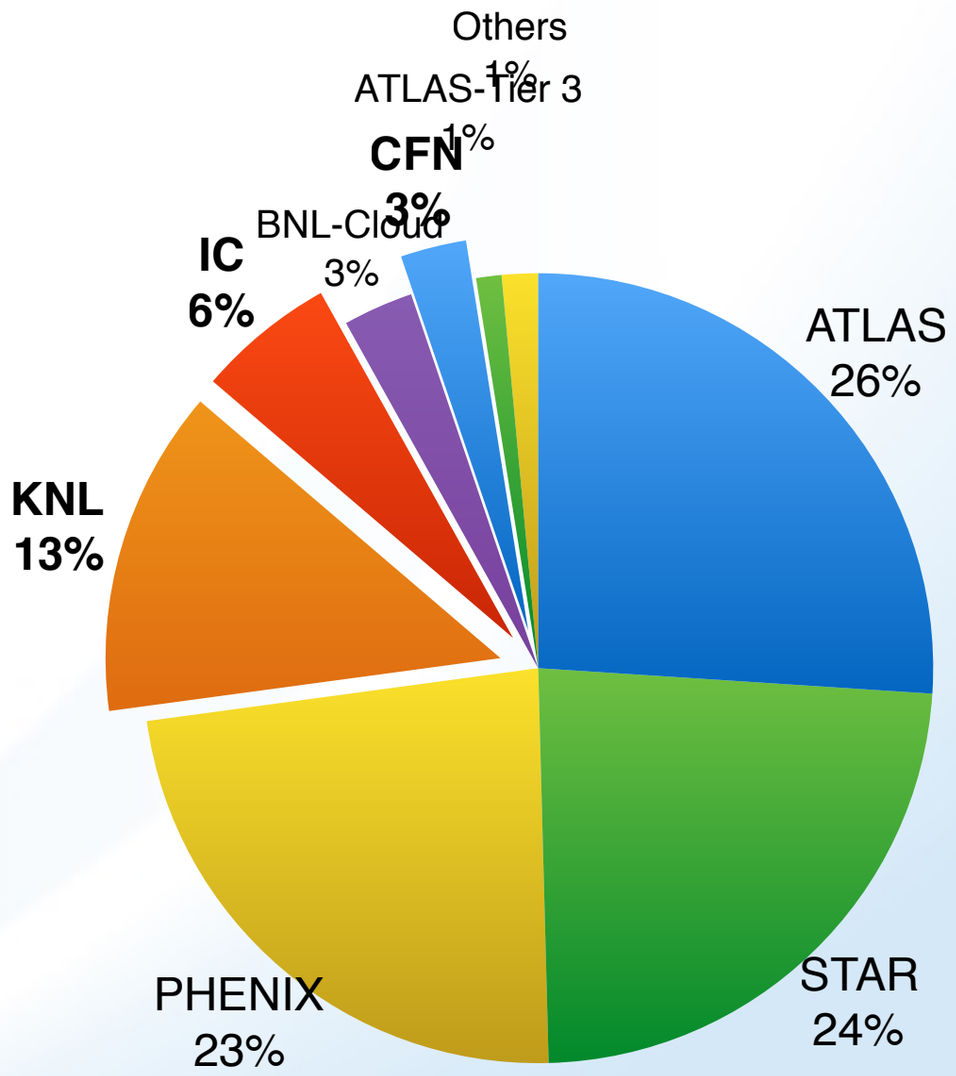
SDCC evolution over the next months

- **Institutional Cluster (IC)**
  - Fall 2016 : 3’888 cores
    (Intel Xeon + Nvidia K80 GPU + InfiniBand)
  - 2 x mid-2017
- **Intel Knights Landing (KNL) cluster**
  - Fall 2016 : 9’000 cores
    (Intel Xeon Phi + Omni-Path)
  - for RBRC (RIKEN and BNL Research Center)
  - and for CSI partners
- **USQCD Cluster**
  - Configuration TBD

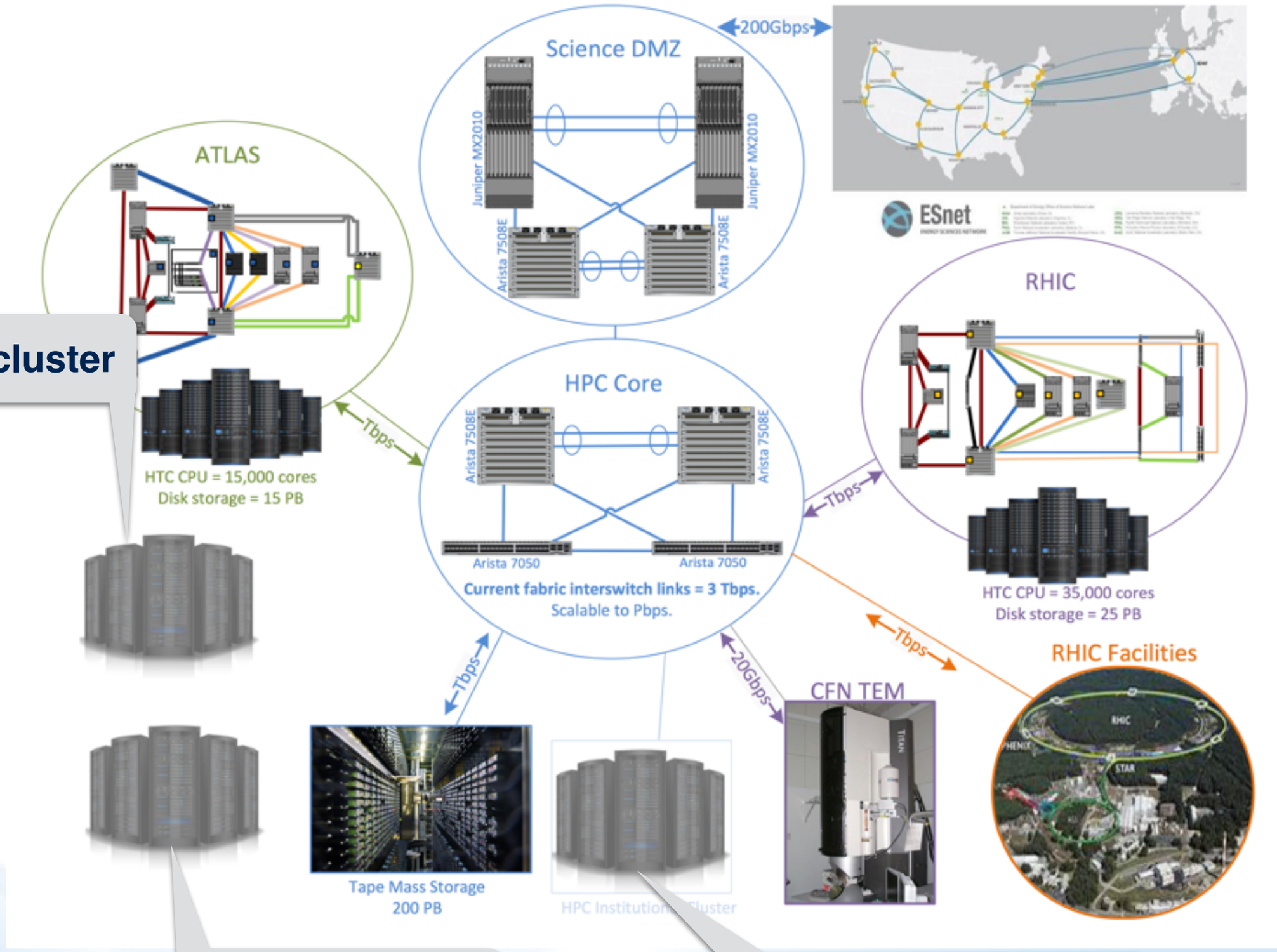
End of 2016

70k cores

HPC : 22%



2017



KNL cluster

USQCD cluster

Institutional Cluster

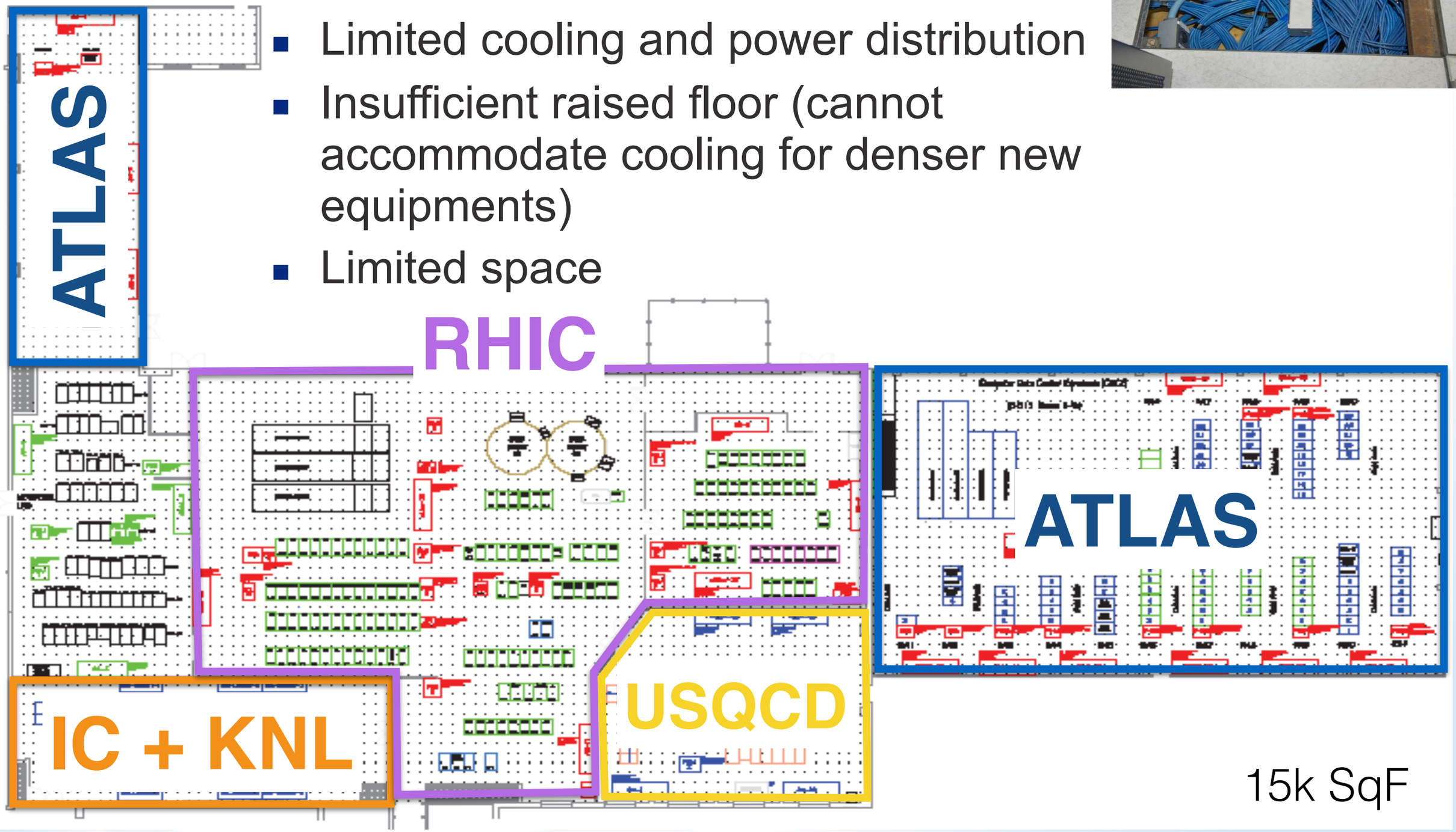
Synergies with BNL Computing Initiative

- CSI is purchasing or complementing purchases in the area of HPC computing (multi-core interconnected nodes)
  - Institutional cluster (Fall 2016, 2x 2017)
  - Knight Landings (KNL) Intel farm (Fall 2016). Initiated by BNL QCD group and RIKEN, CSI doubled the capacity
  - These resources will be made available to RHIC program in opportunistic mode
    - May add 10% to RHIC resources?
    - Issue : manpower to port RHIC codes on KNL?

- Leverage on expertise in data processing & storage technologies developed for RHIC and ATLAS
- Common network, CSI interested in usage of HPSS

Computing room(s) in 2017

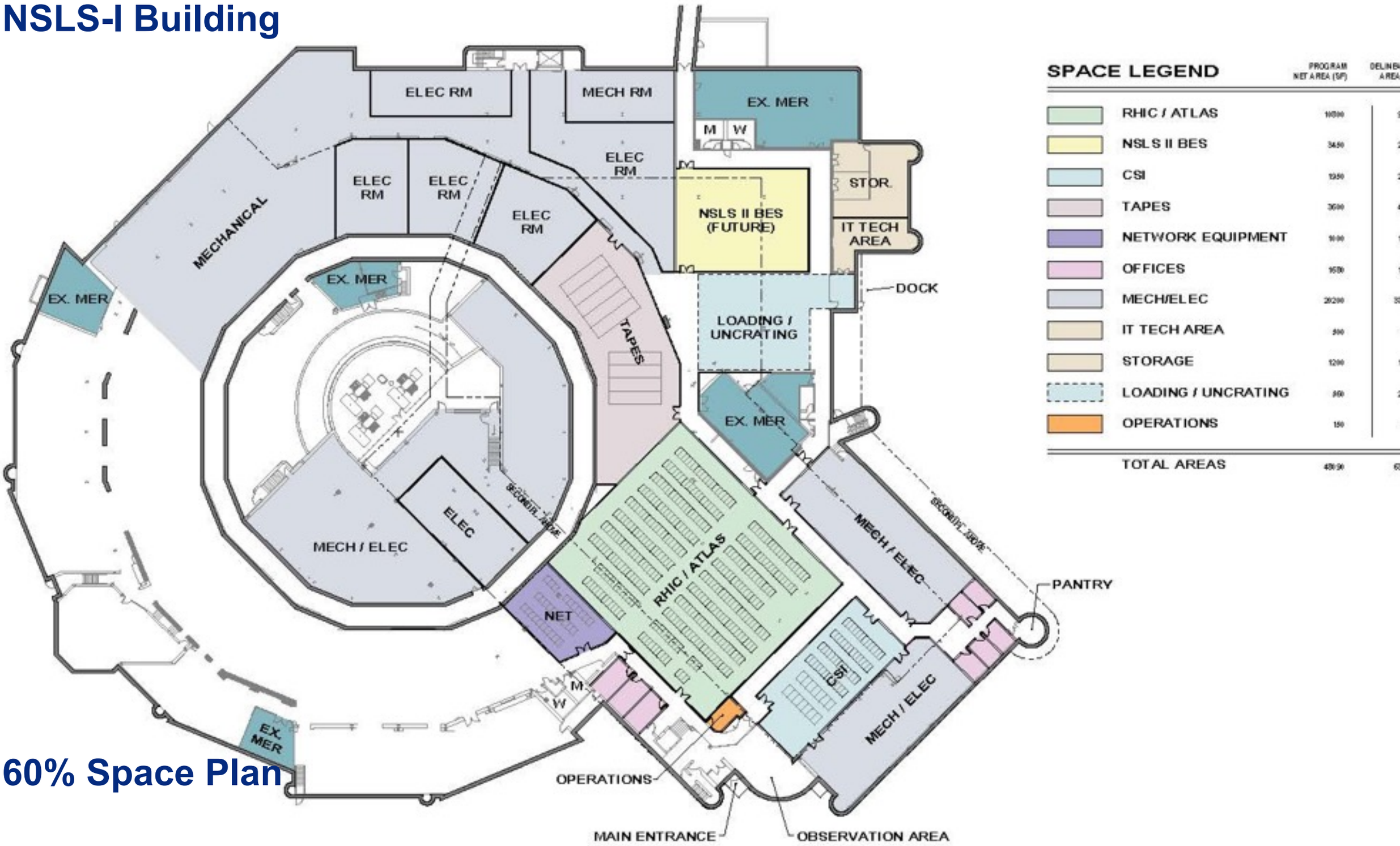
- OLD installations
- Limited cooling and power distribution
- Insufficient raised floor (cannot accommodate cooling for denser new equipments)
- Limited space



New computing room needed

Core Facility Revitalisation – Conceptual Design

NSLS-I Building



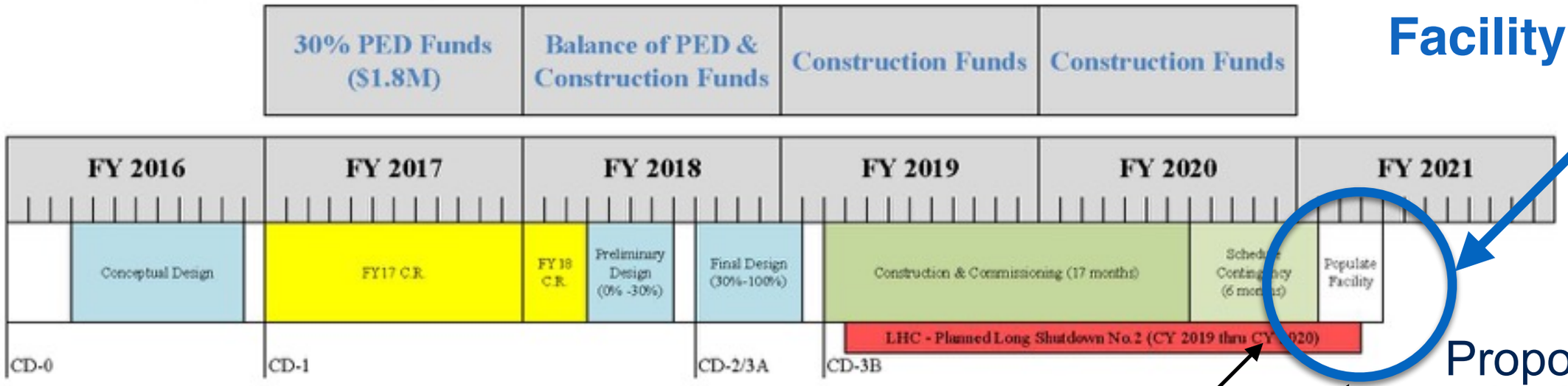
SPACE LEGEND	PROGRAM NET AREA (SF)	DELINEATED AREA (SF)
RHIC / ATLAS	16000	9271
NSLS II BES	3450	2988
CSI	1950	2238
TAPES	3000	4189
NETWORK EQUIPMENT	1000	1186
OFFICES	1500	1186
MECH/ELEC	28200	37584
IT TECH AREA	500	643
STORAGE	1200	1289
LOADING / UNCRATING	800	2879
OPERATIONS	150	132
TOTAL AREAS	48690	63476

60% Space Plan

CD-1 review end of August

CFR – Preliminary Schedule

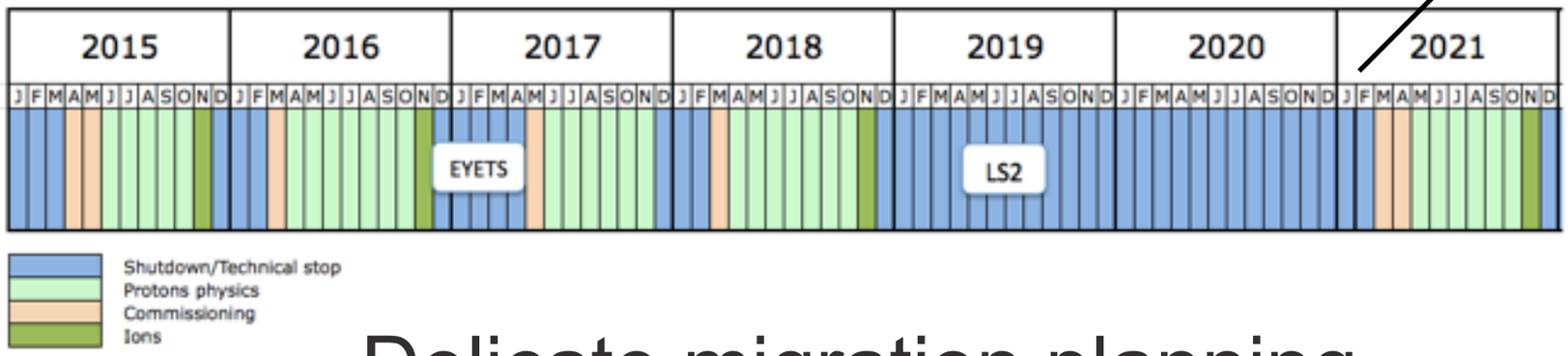
Preliminary CFR Funding Analysis - 1 Yr CR 2017 (Renovation Alternative)



Facility migration

Longer term LHC schedule

The outline LHC schedule out to 2035 presented by Frederick Bordry to the SPC and FC June 2015 can be found here



- Delicate migration planning

Proposed run schedule for RHIC

Years	Beam Species and Energies
2016	High statistics Au+Au d+Au energy scan
2017	High statistics Pol. p+p at 510 GeV
2018	^{96}Zr+^{96}Ru isobar run
2019-20	7.7-20 GeV Au+Au (BES-2)
2021	No Run ?
2022-23	200 GeV Au+Au with upgraded detectors Pol. p+p, p+Au at 200 GeV
2024---	Program TBD

CFR Design – An Incremental Approach

- **Power**
  - Day-one capability (2021) – 2.4 MW IT power (dedicated computing power). This is approximately double current RACF IT power.
  - Provide provision for future1.2 MW IT power increments to 6MW Max.
- **Cooling**
  - Day-one cooling capability to support 2.4 MW IT power
  - Provide provision for future 1.2 MW IT power deployments
- **Space**
  - Day-one - Accommodate approximately 33% footprint expansion (Racks) within defined spaces.
  - Day-one - Accommodate approximately 3,500 SF additional, unassigned space.
  - Provide opportunity for future (long term) growth within the balance of the 725 facility. Both computing and offices.

Summary

- RCF performed remarkably well during Run 16

- Needs for replacement of old hardware, new tape generation & resources needs for 2017 and beyond
  - difficult with level of current budget

- Plan being developed for migrating facility to state of the art computing room in 2021